Big Data: A Framework For Research

Tadhg NAGLE[a,1] and David SAMMON[a]
[a] *Business Information Systems, University College Cork*

Abstract. *Big Data* is not the first and most definitely not the last new term that the IT industry is going to coin in order to drive interest and investment in new technology. Moreover, with these new terms, an opportunity is afforded for the research community to objectively understand the impact (or lack thereof) on organizations and decision makers. This paper provides a high-level framework to guide researchers in the area of *Big Data* through a conceptualization of the Information Supply Chain. The Information Supply Chain can be used as a scoping device for researchers in positioning their work but also as a tool to enable stronger objectivity and prevent an automatic resistance or acceptance of the new term/trend.

Keywords. *Big Data*, Information Supply Chain,

1. Introduction

Big Data, has come to denote a need for an increased technology capability to deal with the large amount of digital information being created each day. This technology centric view is based on the increasing growth of digital information, now estimated to be at 2.5 exabytes per day arising from technological advances such as the 'internet of things' [1]. While traditional technologies are seen as a bottleneck in processing these large data sets [2], new technologies such as Hadoop have provided a platform to store and process such data sets. Yet, while the focus of these technologies are quite immature, being built for very specific tasks, they do point to a new breed of data technologies that inherently deal with the data challenges of today's world. However, what is *Big* (a large volume of) *Data* for an SME is not *Big Data* for a multinational, or indeed what is *Big Data* now is not necessarily going to be *Big Data* in the future. As a result a more objective description of *Big Data* is data that pushes the limits of common technology available at that time. For instance, a gigabyte PowerPoint presentation, a terabyte image or a petabyte movie are all instances of *Big Data* that cannot be adequately managed by applications that currently use them [3]. Moreover, big has evolved from not only meaning volume, but also variety, velocity, extending to validity, veracity, visibility and value [4]. All this highlights a need for technology to not only handle large volumes of data but also deal with practicalities of managing data in organizations.

The perspective taken by this paper is that *Big Data* is a term that describes an evolution of the data capability within organizations. Shifting from a purely technology

[1] Corresponding Author.

centric view this perspective encapsulates the increasing ability of organizations to extract value from the data they possess or have access to. This socio-technical view is not tied to a particular technology, data volume or variety, but does indicate a greater realization of the value of organizational data and greater ability to appropriate that value. In a way, *Big Data* marks a new era or watershed moment, in which the impact of data on our businesses and lives is beginning to accelerate exponentially. As already mentioned we are creating 2.5 exabytes a day [1] and since the 1980's our capability to communicate and store information has doubled every three years [5]. This points to the notion that we are datafying [6] our world and are at a stage where the impact of this datafication process is accelerating change at all levels of society and business[6]. For instance, by datafying the collisions of sub atomic particles the Large Hadron Collider has found the Higgs Boson (popularly known as the god particle). From a more personal perspective the term 'quantified self' describes a movement of datafying one's own daily life (eg amount of steps taken, emails read, hours slept, blood pressure…etc) through new personal devices to make more informed decisions and live a longer happier life [7]. The same impact is also evident for organizations, be it an SME or multinational. For a multinational such as Tesco, an increased data capability has allowed them to identify a saving of €20 million in cooling costs through analyzing refrigeration data within their stores [8]. For the LAPD an increased data capability has translated as analyzing recent criminal events through a specific algorithm developed for earthquakes and aftershocks, to predict places where crime will happen in the near future [9]. For an SME the increased data capability translates as being able to carry out experiments on their website design (using a web analytics platform) on finding how best to configure the site and encourage customers to stay longer.

To fully appreciate this increased data capability that *Big Data* incorporates, this paper outlines the Information Supply Chain, a framework that enables researchers to dissect the capability and provide direction in positioning their research.

2. Information Supply Chain

Describing the sequential flow of data, the Information Supply Chain is developed from an Information Systems and socio-technical perspective, incorporating people, process and technology as key components; a key differentiator from other early *Big Data* research frameworks. For instance, Chen *et al.* [10] published a rigorous account of research in the domain of Business Intelligence and Analytics, highlighting *Big Data* as an emerging area for research. The paper also went on to identify potential areas for research within *Big Data*, but did so from a very technological and analytical centric focus, with spatial mining and Hadoop being two examples from the list [10]. While the Information Supply Chain is technology agnostic it does not diminish the importance of technology. It does however; centralize data as the key focus and as a result reduces the possibility of researching a technology without incorporating the overall impact it may have on an organizations people and processes. A more recent guide for future research describes a very similar framework to the Information Supply Chain but again fails to incorporate significant capacity for the inclusion of the social aspects of *Big Data* research [11]. Instead, it provides a technology centric focus and its impact on data analytics. Another defining characteristic of the Information Supply Chain as a framework is that it is applicable across a number of units of analysis. The

framework can just as easily be applied to a process, decision, organization or even business unit.

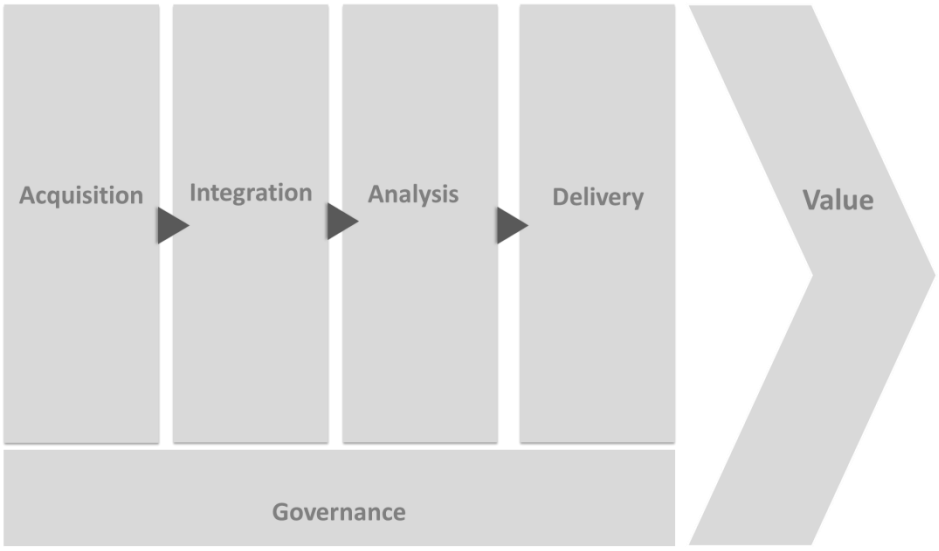

Figure 1. Information Supply Chain

Depicted in Figure 1, the information supply chain incorporates six sections, namely:

(i) Acquisition – associated with the gathering of data and the starting point of the chain.
(ii) Integration – the consolidation and management of the data (both logical and physical) to provide a solid basis for analysis
(iii) Analysis – the application of algorithms and processes to provide insight into the data.
(iv) Delivery – supplying the analysis in a format that provides ease of interpretation and communication of analysis results.
(v) Governance – underpins the chain by emphasizing the need for data quality, integrity, privacy and security, through structured programs and policies.
(vi) Value – the measurable output from the supply chain. Provides an output focus throughout the value chain ensuring a clear objective is defined along with key measures to determine if the object was achieved.

2.1. Acquisition

One notion that *Big Data* has brought into the consciousness of organizations, through the use of success stories and anecdotes, is the ability to mine value from the data within organizational repositories. Moreover, current estimates have highlighted that 80% of data is unstructured [12], which was traditionally overlooked due to the lack of an analytical capability for this type of data. However, now in the era of *Big Data* this 80% has become a landscape of unexplored potential, with the belief that riches are

bubbling underneath the surface, requiring the touch of the right analytical tool. Insight into this potential can be seen in social media applications that have essentially datafied our social relationships as well as (but not limited to) our conversations, likes, dislikes and sentiments towards commercial brands. However, as of yet we are still at a very early stage of understanding unstructured data in itself as a source or starting point for better decision making. Also, with all the emphasis on unstructured data we should not disregard the role of structured data and all the different data structures in-between (eg semantic data structured in triples).

2.2. Integration

The second section in the framework describes the integration and management of data. Questions such as how will organizations deal with managing the increased complexity in integrating additional sources (with different structures) to provide a foundation for improved insight? Questions such as this become more poignant when research shows that organizations are already struggling to integrate and manage their existing data sets. Should these new or previously unexplored data sources be integrated or should they be kept separate in *Big Data* sandboxes/silos. It will be interesting to note the traditional role of the data modeler in the era of *Big Data*. Data modeling essentially creates a blueprint for capturing and analyzing data, ultimately supporting if not defining organizational processes. Such an important role cannot be overlooked or substituted with the promise of new technologies which consequently are not currently configured with built in models.

2.3. Analysis

As already mentioned, the analytical side of *Big Data* has garnered the bulk of research attention. Such attention has also spilled out to the business domain where the title data scientist has been labeled the sexiest job of the 21st century [13]. Yet, there has been very little time given to the need and role of the data savvy manager, which will outnumber data scientists by 10:1 [14]. What good are new technologies and algorithms if business managers are unable to communicate or understand their *Big Data* developer counterparts? Ultimately, responsibility lies with these business managers to identify and steer successful *Big Data* projects and without the development of a common understanding between these two stakeholders, the success rate is likely to be on the low side. This bridging point between analytics and business professionals is a critical success factor, as proven in the IT domain [15, 16] and a definite area for much needed research as organizations aim to develop their data capability.

2.4. Delivery

Delivery is the stage where the output of the previous stages is consumed by the intended audience of the Information Supply Chain. Human computer interaction is already a domain that facilitates the efficient interplay between people and technical artefacts and is set to become an important catalyst in developing new methods that

enable analyst's present/report their results in a manner the intended audience will best understand. Just like Data Scientists, Data Visualizers are also in high demand enabling data consumers absorb the full value of the analysis. Outside of reports there is the delivery of data in a form that is not tightly bounded in a report but in streamlined datasets for end user experimentation and hypothesis testing. However, this format of delivery and supporting technologies should not be perceived as a substitute for other stages in the Information Supply Chain; an error that is still being felt by organizations use of spreadsheets outside of the delivery stage [17].

2.5. Governance

The increasing emphasis on the importance of data leads to the logical assumption for the need of an increase in the importance of data governance. Yet, unfortunately only 10% of organizations have implemented data governance initiatives in the first place [18]. Without strong data governance, organizations may mindlessly jump on the *Big Data* bandwagon (a characteristic of the IT domain) [19]. For instance, the current rhetoric around *Big Data* presumes all data as an asset. Good data governance research may challenge that assumption and as a result view certain datasets as possible liabilities, which in turn may provide guidance for focused *Big Data* implementations that are cognizant of any incumbent risks. In essence, organizations need to know their data maturity and readiness for *Big Data* to prevent inappropriate analysis of data leading to incorrect decisions or even a marginalization of everything that is not digital.

2.6. Value

Ultimately the success of any *Big Data* project will depend on the value it has created. Yet, there is the feeling that in the majority of cases the actual value created by new *Big Data* technologies is unclear at the time of procurement, putting the onus on customers to figure it out thereafter. As a comparison, research in working out the business value of IT has been a difficult task to say the least and has only recently proved definitively that IT does add value [20]. The complexity in linking back value creation to data projects is also set to be difficult and as a result more difficult to justify *Big Data* initiatives in the first place. *Big Data* needs to move beyond idiosyncratic anecdotes of success, which offer very little value in transferring the success to differing cases. There are opportunities for research in developing value creation patterns that can be achieved by a wide range of organizations.

3. Conclusion

The aim of this paper has been to outline the early skeleton of a framework in which researchers can identify areas of study within the domain of *Big Data.* We feel the use of the Information Supply Chain provides a lens that uniquely emphasizes the socio-technical impact that *Big Data* is going to have on organizations and society. While still in the early stages of the development we also see practical as well as academic uses for the framework. For instance, similar to what the Business Model Canvas has

done for organizations by enabling them to visually represent and reflect on their business model, the Information Supply Chain enables researchers and practitioners alike to create a mental model of an organizations data capability and highlight potential future opportunities.

References

[1] IBM, What is Big Data?, in, http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html, 2013.
[2] E.A. Marks, B. Lozano, Executive's guide to cloud computing, Wiley. com, 2010.
[3] EMC, Big Ideas: How Big is Big Data?, in: P. Florissi (Ed.), http://www.youtube.com/watch?v=eEpxN0htRKI, 2012.
[4] R. Livingstone, The 7 Vs of Big Data, in, http://rob-livingstone.com/2013/06/big-data-or-black-hole/, 2013.
[5] M. Hilbert, P. López, Global telecommunication and storage capacity doubling every three years The World's Technological Capacity to Store, Communicate, and Compute Information, Science, 332 (2011) 60-65.
[6] M. Lycett, 'Datafication': making sense of (big) data in a complex world, European Journal of Information Systems, 22 (2013) 381-386.
[7] Wikipedia, Quantified Self, in, http://en.wikipedia.org/wiki/Quantified_Self, 2013.
[8] B. Goodwin, Tesco uses big data to cut cooling costs by up to €20m, in: ComputerWeekly.com, http://www.computerweekly.com/news/2240184482/Tesco-uses-big-data-to-cut-cooling-costs-by-up-to-20m, 2013.
[9] A. Mendelson, Can LAPD Anticipate Crime With 'Predictive Policing'? - See more at: http://www.californiareport.org/archive/R201309061630/b#sthash.OpksGgr3.dpuf, in: The California Report, http://www.californiareport.org/archive/R201309061630/b, 2013.
[10] H. Chen, R.H. Chiang, V.C. Storey, Business Intelligence and Analytics: From Big Data to Big Impact, MIS Quarterly, 36 (2012) 1165-1188.
[11] D. Agrawal, S. Das, A. El Abbadi, Big data and cloud computing: current state and future opportunities, in: Proceedings of the 14th International Conference on Extending Database Technology, ACM, 2011, pp. 530-533.
[12] Gartner, Top 10 Strategic Technology Trends for 2012, (2011).
[13] T.H. Davenport, D. Patil, Data scientist: the sexiest job of the 21st century, Harv. Bus. Rev., 90(2012) 70-77.
[14] McKinsey, Big data: The next frontier for competition, in, http://www.mckinsey.com/features/big_data, 2013.
[15] J.C. Henderson, N. Venkatraman, Strategic alignment; a model for organizational transformation via information technology, in: T.J. Allen, M.S. Scott Morton (Eds.) Information Technology and the Corporation of the 1990 's, Oxford University Press, Oxford, 1994, pp. 202-220.
[16] J. Luftman, T. Brier, Achieving and sustaining business-IT alignment, CALIFORNIA MANAGEMENT REVIEW, 42 (1999) 109-122.
[17] P. O'Beirne, Spreadsheet mistakes, in, http://www.eusprig.org/horror-stories.htm, EuSpRIG 2004.
[18] T. Nagle, D. Sammon, Fast and Flexible: Exploring Agile Data Analytics, Cutter Benchmark Review, 13 (2013) 5-9.
[19] C.M. Fiol, E.J. O'Connor, Waking up! Mindfulness in the face of bandwagons, Academy of Management Review, 28 (2003) 54-70.
[20] S. Mithas, A. Tafti, I. Bardhan, J. Mein Goh, INFORMATION TECHNOLOGY AND FIRM PROFITABILITY: MECHANISMS AND EMPIRICAL EVIDENCE, Mis Quarterly, 36 (2012) 205-224.